Radiation Protection Dosimetry (2010), Vol. 138, No. 4, pp. 363–375
Advance Access publication 26 November 2009
doi:10.1093/rpd/ncp268

POISSON MIXTURE MODEL FOR MEASUREMENTS USING COUNTING

Guthrie Miller[1,*], Alan Justus[1], Vadim Vostrotin[2], Donald Dry[1] and Luiz Bertelli[1]
[1]Los Alamos National Laboratory, Los Alamos, NM, USA
[2]Southern Urals Biophysics Institute, Ozyorsk, Chelyabinsk, Russian Federation

Received May 27 2009, revised September 9 2009, accepted October 26 2009

Starting with the basic Poisson statistical model of a counting measurement process, 'extraPoisson' variance or 'overdispersion' are included by assuming that the Poisson parameter representing the mean number of counts itself comes from another distribution. The Poisson parameter is assumed to be given by the quantity of interest in the inference process multiplied by a lognormally distributed normalising coefficient plus an additional lognormal background that might be correlated with the normalising coefficient (shared uncertainty). The example of lognormal environmental background in uranium urine data is discussed. An additional uncorrelated background is also included. The uncorrelated background is estimated from a background count measurement using Bayesian arguments. The rather complex formulas are validated using Monte Carlo. An analytical expression is obtained for the probability distribution of gross counts coming from the uncorrelated background, which allows straightforward calculation of a classical decision level in the form of a gross-count alarm point with a desired false-positive rate. The main purpose of this paper is to derive formulas for exact likelihood calculations in the case of various kinds of backgrounds.

INTRODUCTION

In making statistical inferences that allow quantitative estimates of uncertainty, one needs a statistical model of the measurement process. As an example, in internal dosimetry the quantitative estimation of the uncertainty of an internally deposited radiation dose is a current topic of great interest[(1)]. The underlying statistical model for counting measurements is the Poisson model; however, it is well recognised that 'extraPoisson variance' or 'overdispersion' may exist caused by the Poisson parameter representing the mean expected number of counts itself coming from some other distribution. This present work paper revisits and revises some previous work[(2–5)], where the other distribution is assumed to be lognormal. The usual treatment of Poisson overdispersion assumes that the other distribution is a Gamma distribution, in which case the resulting mixture distribution is again a Gamma distribution. However, the lognormal seems to agree with data and is qualitatively different from the Gamma distribution in that it can have at the same time a finite mean and an unlimited standard deviation over mean ratio.

The issue of greatest interest in this present work is the question of additional backgrounds. These backgrounds can originate in different places in the measurement process, and the detailed statistical modelling is different. The true value of the Poisson mean counts parameter is assumed equal the true value of the quantity of interest (e.g. the urine excretion rate) times a normalisation coefficient plus a background. While the normalisation coefficient and background are assumed to be lognormal and possibly correlated, another, so-called type-0 background that is not correlated with the normalisation coefficient is also assumed to be present. The Poisson parameter of the type-0 background is estimated from the Bayesian posterior distribution after a background measurement, assuming a Gamma prior on the true type-0 background counting rate.

The main purpose of this paper is to derive formulas for exact likelihood calculations in the case of various kinds of backgrounds. The likelihood function is central to statistical inference from data, as in internal dosimetry. The intent of the paper is to simplify the derivations and perform numerical checks using Monte Carlo. Also, exact analytical expressions for the variance of the data are derived that allow calculation of χ^2 for statistical self-consistency checking. The type-0 background discussed here is the basis of most treatments of the classical decision level (minimum false-positive rates). More discussion and literature references are given in the last section of this paper.

Theoretical model

The counting measurement result is some positive integer N, which is assumed to have a Poisson

*Corresponding author: guthrie@lanl.gov

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from rpd.oxfordjournals.org by guest on October 4, 2010

distribution with mean value denoted by μ. The probability of measuring N counts is then given by

$$P(N) = \frac{\mu^N e^{-\mu}}{N!}. \tag{1}$$

If μ is a known constant, this model describes simple counting measurements. In simple counting the variance of the number of counts is given by

$$\begin{aligned} \mathrm{Var}(N) &= E(N^2) - (E(N))^2 = \mu \\ &= E(N), \end{aligned} \tag{2}$$

where

$$\begin{aligned} E(N) &= \sum_{N=0}^{\infty} NP(N) = \mu \\ E(N^2) &= \sum_{N=0}^{\infty} N^2 P(N) = \mu^2 + \mu. \end{aligned} \tag{3}$$

Here, the situation where μ is given by a known distribution rather than being constant is considered. The probability of measuring N counts is then given by

$$P(N) = \int d\mu P(\mu) \frac{\mu^N e^{-\mu}}{N!}. \tag{4}$$

Thus, the distribution of the measured data is a continuous mixture of Poisson distributions with different values of μ.

It is helpful to note two limiting cases of the probability distribution of the measured number of counts from Eq. (4). When the average value of μ is small, the composite distribution is Poisson with mean value equal to the average value of μ; however, at the other extreme the distribution of N given μ becomes a narrow Gaussian, and the distribution of N approaches the distribution of μ.

Using Eq. (4) and interchanging summation and integration, the mean and variance of N are given by

$$\begin{aligned} E(N) &= \sum_{N=0}^{\infty} NP(N) = \int d\mu P(\mu)(\mu) = E(\mu) \\ E(N^2) &= \sum_{N=0}^{\infty} N^2 P(N) = \int d\mu P(\mu)(\mu^2 + \mu) \\ &= E(\mu^2) + E(\mu) \end{aligned} \tag{5}$$

and therefore the variance is given by

$$\begin{aligned} \mathrm{Var}(N) &= E(N^2) - (E(N))^2 \\ &= E(\mu) + \mathrm{Var}(\mu) \\ &= E(N) + \mathrm{Var}(\mu). \end{aligned} \tag{6}$$

Contrasting Eq. (2) for simple counting with Eq. (6), one sees that the variance of N is increased by the variance of μ.

In the situation under consideration the true value of the quantity of interest is denoted by ψ (for example, 24-h urine excretion) and is related to μ using the following formula involving a normalisation coefficient A and a background B:

$$\mu = A\psi + B. \tag{7}$$

In what follows the value of ψ is assumed to be given, as is appropriate for the problem of inferring ψ from the data. The units of μ and B are counts, while ψ has some physical units (e.g. Bq), and A has units of counts per physical unit.

The likelihood function is defined as $P(N|\psi)$ considered as a function of ψ. It is central for statistical inference (of ψ) from data (N), and it is important to calculate it as accurately as possible. The likelihood function is given by (up to an unimportant constant multiplicative factor)

$$\begin{aligned} L(\psi) &\propto P(N|\psi) \propto \int (\mu(\psi))^N e^{-\mu(\psi)} P(A,B) \mathrm{dAdB} \\ \mu(\psi) &= A\psi + B, \end{aligned} \tag{8}$$

In order to calculate the likelihood function one must properly take into account the correlations between A and B.

The means (expectation values) of μ and μ^2 are given by

$$\begin{aligned} E(\mu) &= \int (A\psi + B)\, P(A,B) dA dB \\ &= \psi E(A) + E(B) \\ E(\mu^2) &= \int (A\psi + B)^2 P(A,B) dA dB \\ &= \psi^2 E(A^2) + 2\psi E(AB) + E(B^2). \end{aligned} \tag{9}$$

The variance of μ is therefore

$$\begin{aligned} \mathrm{Var}(\mu) &= E(\mu^2) - (E(\mu))^2 \\ &= \psi^2 \mathrm{Var}(A) + 2\psi \mathrm{Cov}(A,B) \\ &\quad + \mathrm{Var}(B), \end{aligned} \tag{10}$$

where the covariance is defined as

$$\mathrm{Cov}(A,B) = E(AB) - E(A)E(B). \tag{11}$$

It is conventional to define the measurement value of the quantity of interest, although using exact likelihood calculations this turns out not to be essential. The measurement value will be denoted by y (measurement value denoted by Roman character y, true value by Greek character ψ), and it is given by

Downloaded from rpd.oxfordjournals.org by guest on October 4, 2010

the formula

$$y = \frac{N - B_0}{A_0}, \tag{12}$$

where B_0 is assumed to be the mean value of B and A_0 is chosen to be the median value of A, which makes the lognormal limit have a simple form as will be discussed later on. The mean and variance of y are given by

$$E(y) = \frac{\psi E(A) + E(B) - B_0}{A_0}$$

$$\mathrm{Var}(y) = \frac{E(B) + \mathrm{Var}(B) + \psi(E(A) + 2\mathrm{Cov}(A,B)) + \psi^2 \mathrm{Var}(A)}{A_0^2}. \tag{13}$$

Background correlations: four types of background in radiochemical urine measurement

A measurement of urine to determine the activity excreted in a 24 h period is assumed. In this case, the normalisation coefficient A is given by

$$A = \Delta t_x \varepsilon_r \varepsilon_c \Delta t_c, \tag{14}$$

where Δt_x is the excretion time associated with the sample in units of 24 h, ε_r is the efficiency of chemical recovery, ε_c is the counting efficiency and Δt_c is the count time. These quantities are assumed to have independent distributions. In practice, Δt_x might be obtained for each sample from the collection time, if a true 24-h collection protocol is followed, or from a volume and specific gravity measurement of the urine sample[6]. The efficiency of chemical recovery ε_r might be obtained for each sample from a measurement of tracer recovery. The uncertainty, given the measurements (e.g. of volume and specific gravity or tracer recovery), is what is of concern here. The counting efficiency ε_c is usually measured infrequently, and this measured value and an estimate of its uncertainty is used for many samples. The count time is usually known very accurately.

Four different types of background B can be imagined depending on where the contamination that causes the background occurs, as shown in Table 1. The actual background is assumed to be the sum of a type-0 background and a background of types 1, 2, or 3. The type-0 background is distinguished by being independent of the normalisation coefficient (no covariance).

The quantity b is the random variable representing the background, ε_c the random variable representing counting efficiency, etc. The quantities with a tilde above, for example $\hat{\varepsilon}_c$, are median values used for normalisation purposes. The point is that, depending upon where the background originates, it has additional random variation caused by the additional efficiency factors. Thus, the type-1 background is proportional to $b\varepsilon_c$ and has random variation because of the variation of the background itself and also because of variation of the counting efficiency.

Type-0 background

It is conventional in considering Poisson mixture distributions to assume the distribution of the Poisson parameter to be given by a Gamma distribution, because of the closed form expressions that result. The true background counting rate λ is then assumed to have the following distribution:

$$P(\lambda)\mathrm{d}\lambda = \mathrm{Gamma}(\lambda, \alpha, \beta)\mathrm{d}\lambda \equiv \frac{\beta^\alpha}{(\alpha-1)!}\lambda^{\alpha-1}\mathrm{e}^{-\beta\lambda}\mathrm{d}\lambda \propto \lambda^{\alpha-1}\mathrm{e}^{-\beta\lambda}\mathrm{d}\lambda, \tag{15}$$

where α and β are parameters. The expectation, mode and variance are given by

$$E(\lambda) = \frac{\alpha}{\beta}$$

$$\lambda_{\mathrm{mode}} = \frac{\alpha - 1}{\beta}.$$

Table 1. Definitions of four different types of background for a radiochemical measurement.

Type	Location	B	Cov (A,B)
0	On the detector itself	b	0
1	On the sample (for example contamination in the last stages of the chemistry)	$b\frac{\varepsilon_c}{\hat{\varepsilon}_c}$	$E(b)E(\Delta t_x)E(\varepsilon_r)\frac{\mathrm{Var}(\varepsilon_c)}{\hat{\varepsilon}_c}\Delta t_c$
2	In the measured aliquot before chemistry (for example, tracer contamination)	$b\frac{\varepsilon_r\varepsilon_c}{\hat{\varepsilon}_r\hat{\varepsilon}_c}$	$E(b)E(\Delta t_x)\frac{\mathrm{Var}(\varepsilon_r\varepsilon_c)}{\hat{\varepsilon}_r\hat{\varepsilon}_c}\Delta t_c$
3	In the person's urine (for example, because of environmental intakes)	$b\frac{\Delta t_x\varepsilon_r\varepsilon_c}{\hat{\varepsilon}_x\hat{\varepsilon}_r\hat{\varepsilon}_c}$	$E(b)\frac{\mathrm{Var}(\Delta t_x\varepsilon_r\varepsilon_c)}{\hat{\varepsilon}_x\hat{\varepsilon}_r\hat{\varepsilon}_c}\Delta t_c$

Downloaded from rpd.oxfordjournals.org by guest on October 4, 2010

$$\mathrm{Var}(\lambda) = \frac{\alpha}{\beta^2}$$

Note that the choice of parameters $\alpha = 1$, $\beta \to 0$ leads to a uniform (flat) distribution. Because $\mu \equiv \lambda T$,

$$\begin{aligned} E(\mu) &= TE(\lambda) = \frac{\alpha}{\beta'} \\ \mathrm{Var}(\mu) &= T^2\mathrm{Var}(\lambda) = \frac{\alpha}{\beta'^2}, \end{aligned} \tag{16}$$

where $\beta' = \beta/T$.

The type-0 background B_0 in Eq. (12) is the mean value from Eq. (16), and the definition of y becomes

$$\begin{aligned} y &= \frac{N - E(\mu) - B_0}{A_0} \\ &= \frac{N - \alpha/\beta' - B_0}{A_0}, \end{aligned} \tag{17}$$

where B_0 above refers to the additional background of types 1, 2 or 3.

From Eq. (6) the variance of y coming from the type-0 background is

$$\begin{aligned} \mathrm{Var}(y) &= \frac{E(\mu) + \mathrm{Var}(\mu)}{A_0^2} \\ &= \frac{1}{A_0^2}\frac{\alpha}{\beta'}\left(1 + \frac{1}{\beta'}\right). \end{aligned} \tag{18}$$

An uncorrelated type-0 background is simply additive (in variance) to backgrounds of the other types.

The question then becomes: how are the parameters α and β' to be determined?

Imagine a measurement of the type-0 background by counting a blank sample for time T_B, which is $R = T_B/T$ times longer than normal. The number of detected background counts is denoted by N_B. The Bayesian interpretation of this measurement involves a prior for the background counting rate λ, which is assumed to be a Gamma function with parameters α_0 and β_0. By Bayes theorem, the posterior probability of λ is also a Gamma function (see Appendix A):

$$\begin{aligned} P(\lambda|N_B) &\propto P(N_B|\lambda)P(\lambda) \propto (\lambda T_B)^{N_B} e^{-\lambda T_B} \lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda} \\ &\propto \lambda^{N_B + \alpha_0 - 1} e^{-(T_B + \beta_0)\lambda} \\ &\propto \mathrm{Gamma}(\lambda, N_B + \alpha_0, T_B + \beta_0). \end{aligned} \tag{19}$$

In this way, the determination $\alpha = N_B + \alpha_0$ and $\beta = T_B + \beta_0$ have been made so that $\beta' \equiv \beta/T = (T_B + \beta_0)/T = R + \beta_0'$. This Bayesian method of determining α and β will be used in the remainder of this paper.

In the conventional method of treating uncertainty of background, the net counts are $N - N_B/R$, and the variance of net counts is calculated as $N + N_B/R^2$, using $\mathrm{Var}(N) = N$ for a Poisson deviate N. When there is no true amount (no radioactivity in the sample), the gross counts would be given by N_B/R, and the variance of net counts becomes $N_B/R + N_B/R^2$. One can easily go back and forth between this characterisation of background and the conventional one by using the replacements $N_B + \alpha_0 \leftrightarrow N_B$ and $R + \beta_0' \leftrightarrow R$, which might be helpful in adapting existing software. It is interesting that in the conventional analysis, part of the variance of net counts comes from the variance of background counts, while in the analysis here, the number of background counts varies randomly; however, after a background measurement, the number of background counts (the number actually measured) is treated as a given, without random variation. In the analysis here, the variance of net counts comes entirely from the variance of gross counts.

It is also useful to consider the probability distribution of measured background counts, because this is an observable of interest and, using the definition of the normalised Gamma function, it can be evaluated in closed form as follows:

$$\begin{aligned} P(N|\alpha, \beta) &= \int d\lambda \, \frac{(\lambda T)^N}{N!} e^{-\lambda T} \mathrm{Gamma}(\lambda, \alpha, \beta) \\ &= \frac{1}{(\alpha - 1)!}\left(\frac{\beta'}{1 + \beta'}\right)^{N_B + \alpha} \\ &\quad \frac{(N + \alpha - 1)!}{N!}\left(\frac{1}{1 + \beta'}\right)^N. \end{aligned} \tag{20}$$

An example of the use of this expression to determine α and β from an empirical cumulative distribution of N is given in Appendix A.

Lognormal distributions

At this point, a lognormal form is assumed for the distributions A and B for background types 1, 2 and 3. The lognormal assumption seems a good choice for two reasons: experimentally in agreeing with data[7] and also theoretically in that the product of independent lognormal random variables is again lognormal, allowing the covariance terms between the background and the normalisation coefficient to be evaluated with closed form expressions.

Downloaded from rpd.oxfordjournals.org by guest on October 4, 2010

Table 2. Expressions for terms in formula for Var(y) for different types of background.

Type	Cov term	$\sigma_0^2\hat{A}^2$	$E(B)$
0	0	$\frac{N_B+\alpha_0}{R+\beta_0'}\left(1+\frac{1}{R+\beta_0'}\right)$	$\frac{N_B+\alpha_0}{R+\beta_0'}$
1	$2\hat{b}a_{bc}\,(a_c^2-1)$	$\hat{b}a_{bc}\left(1+\hat{b}a_{bc}(a_{bc}^2-1)\right)$	$\hat{b}a_{bc}$
2	$2\hat{b}a_{brc}\,(a_{rc}^2-1)$	$\hat{b}a_{brc}\left(1+\hat{b}a_{brc}(a_{brc}^2-1)\right)$	$\hat{b}a_{brc}$
3	$2\hat{b}a_{bxrc}\,(a_{xrc}^2-1)$	$\hat{b}a_{bxrc}\left(1+\hat{b}a_{bxrc}(a_{bxrc}^2-1)\right)$	$\hat{b}a_{bxrc}$

The lognormal distribution is given by

$$\mathrm{d}P(f) = \frac{1}{\sqrt{2\pi}S}\exp\left[\frac{-\log(f/\hat{f})^2}{2S^2}\right]d\,\ln(f), \quad (21)$$

where $\hat{f}$ is the median value and S is the logarithmic standard deviation. One can show using elementary integration (starting from basics by completing the square in the exponential) that

$$\int f^p \mathrm{d}P(f) = \hat{f}^p \exp\left(\frac{p^2S^2}{2}\right). \quad (22)$$

Thus,

$$\begin{aligned} E(f) &= \hat{f}\,\mathrm{e}^{S^2/2} \equiv \hat{f}a \\ \mathrm{Var}(f) &= \hat{f}^2a^2\left(a^2-1\right) \end{aligned}. \quad (23)$$

The mean and variance of A are then given by

$$\begin{aligned} E(A) &= \hat{A}a_{xrc} \\ \mathrm{Var}(A) &= \hat{A}^2a_{xrc}^2\left(a_{xrc}^2-1\right) \end{aligned}, \quad (24)$$

where the combined variance from excretion time, chemical recovery and counting efficiency is given by

$$\begin{aligned} S_{xrc}^2 &\equiv S_x^2 + S_r^2 + S_c^2 \\ a_{xrc} &\equiv \mathrm{e}^{S_{xrc}^2/2}, \end{aligned} \quad (25)$$

and the median of A is the product of the medians of the individual factors as follows:

$$\hat{A} = \Delta\hat{t}_x\hat{\varepsilon}_r\hat{\varepsilon}_c\Delta t_c. \quad (26)$$

Choosing A_0 to be the median value and B_0 to be the mean value in Eq. (12) gives

$$E(y) = \psi a_{xrc}\mathrm{Var}(y) = \sigma_0^2 + \frac{\psi a_{xrc}}{\hat{A}}(1+\text{Cov term}) + \psi^2a_{xrc}^2\left(a_{xrc}^2-1\right), \quad (27)$$

where σ_0^2 is Var(y) for a zero true amount in the following sample:

$$\sigma_0^2 = \frac{E(B)+\mathrm{Var}(B)}{\hat{A}^2}. \quad (28)$$

Table 2 shows the covariance term, σ_0^2, and $E(B)$ to allow one to calculate y and Var(y) using Eq. (12), always assuming a type-0 background and with one of the other three types of background.

Monte Carlo validation

To validate these rather complex analytical expressions, a numerical study using Monte Carlo was carried out. The theoretical results were tested by calculating χ^2 defined as follows:

$$\chi_j^2 = \frac{(y_j - E(y))^2}{\mathrm{Var}(y)}, \quad (29)$$

where j denotes the trial and N_{trials} is the number of trials. The expectation value of y is given by $E(y) = \psi E(A)/A_0 = \psi a_{xrc}$. If the theoretical derivation is correct and the Monte Carlo generation is done properly, the average value of χ^2 must approach one for a large number of trials.

The parameter values were as given in Table 3.

The type-0 μ_B is generated from a Gamma distribution Gamma(μ_B/T_B, α_0, β_0), and then this μ_B was used to generate a number of background counts N_B from a Poisson distribution. Given this value of N_B, type-0 μ was again generated from Gamma(μ/T, $N_B+\alpha_0$, $T_B+\beta_0$). Then values of A and B are generated from their lognormal distributions and Eq. (7) is used to calculate the total μ.

Downloaded from rpd.oxfordjournals.org by guest on October 4, 2010

Table 3. Parameter values used in the Monte Carlo validation study.

Parameter	Value
$\hat{A}$	1000 count Bq^{-1}
Δt_c	5333 s
Δt_x	Median = 1, $S = 0.5$
ε_r	Median = 0.75, $S = 0.05$
ε_c	Median = 0.25, $S = 0.05$
Type-0 background	$\alpha_0 = 1$, $\beta_0' = 2$
Other type background	Median:$b = 2$ counts, $S = 1$

The gross counts N are then generated from the Poisson distribution with this value of μ. Equation (17) is used to calculated y_j. In Eq. (29), $\mathrm{Var}(y)$ is calculated from Eq. (27) and Table 2.

Results are given in Table 4 for the four background types ($N_{\mathrm{trials}} = 10^6$). The analytical expressions are confirmed by these results.

Table 4. Results of Monte Carlo validation study ($\chi^2 = 1$ is the expected result).

Type	ψ	Average χ^2
0	0	1.011 ± 0.003
	1	0.994 ± 0.002
	10	0.997 ± 0.003
1	0	0.989 ± 0.008
	1	0.999 ± 0.007
	10	0.993 ± 0.003
2	0	0.998 ± 0.010
	1	1.007 ± 0.008
	10	1.002 ± 0.003
3	0	0.999 ± 0.011
	1	0.972 ± 0.010
	10	1.010 ± 0.006

Generalisation, shared uncertainty of background

In this section, a synthesis of the foregoing is made by introducing the notion of shared uncertainty. The normalisation coefficient A is assumed to have lognormal standard deviation S from all factors. Instead of considering separately backgrounds of types 1, 2 or 3, it is assumed that some portion S_s of S is shared with the background, where $0 \le S_s \le S$. The background is now characterised by its median value in physical units $\hat{B}$, where in terms of the previous notation, $\hat{b}/A_0 = \hat{B}$. The lognormal background is assumed to have logarithmic standard deviation S_b. Then

$$\mathrm{Var}(y) = \sigma_0^2 + \frac{a\psi}{A_0} + 2a\psi\hat{B}a_{\mathrm{bs}}\left(a_s^2 - 1\right) + (a\psi)^2\left(a^2 - 1\right), \tag{30}$$

where

$$\sigma_0^2 = \frac{1}{A_0^2}\frac{(N_B + \alpha_0)}{R + \beta_0'}\left(1 + \frac{1}{R + \beta_0'}\right) + \hat{B}a_{\mathrm{bs}}\left(1 + \hat{B}a_{\mathrm{bs}}\left(a_{\mathrm{bs}}^2 - 1\right)\right)$$

is the variance for the zero true amount and

$$\begin{aligned} a &\equiv \mathrm{e}^{S^2/2} \\ a_s &\equiv \mathrm{e}^{S_s^2/2} \\ a_b &\equiv \mathrm{e}^{S_b^2/2} \\ a_{\mathrm{bs}} &\equiv \mathrm{e}^{S_{\mathrm{bs}}^2/2} = \mathrm{e}^{(S_b^2 + S_s^2)/2}. \end{aligned} \tag{31}$$

This form reproduces the various types of background correlations discussed above.

An alternate approximate form of Eq. (30), which contains the conventional 'counting uncertainty', is obtained by substituting

$$\frac{a\psi}{A_0} = \frac{E(N) - (N_B + \alpha_0)/(R + \beta_0') - A_0\hat{B}a_{\mathrm{bs}}}{A_0^2}, \tag{32}$$

from Eqs (12) and (17) and using $E(N) \cong N$, which gives

$$\begin{aligned} \mathrm{Var}(y) &\cong \sigma_c^2 + \sigma_b^2 + 2a\psi\hat{B}a_{\mathrm{bs}}\left(a_s^2 - 1\right) + (a\psi)^2\left(a^2 - 1\right)\sigma_c^2 \equiv \frac{1}{A_0^2}\left(N + \frac{N_B + \alpha_0}{(R + \beta_0')^2}\right) \\ \sigma_b^2 &\equiv \left(\hat{B}a_{\mathrm{bs}}\right)^2\left(a_{\mathrm{bs}}^2 - 1\right), \end{aligned} \tag{33}$$

where N is the number of detected gross counts and N_B is the number of detected background counts. This form separates out the uncertainty σ_c coming from counting statistics with a constant normalisation coefficient and the variance of the background σ_b. The additional terms result from a non-constant normalisation coefficient.

In terms of shared uncertainty, the likelihood function is given by

$$\begin{aligned} L(\psi) &\propto P(N|\psi) \propto \int \mathrm{d}P(A_n)\,\mathrm{d}P(A_s)\,\mathrm{d}P(b)\,\mathrm{d}P(\mu_0) \\ (\mu(\psi))^N \mathrm{e}^{-\mu(\psi)}\mu(\psi) &= A_0A_s(A_n\psi + b) + \mu_0, \end{aligned} \tag{34}$$

Downloaded from rpd.oxfordjournals.org by guest on October 4, 2010

where the lognormal normalisation coefficient A is factored into two lognormals A_s and A_n with median values 1, where A_s is shared with the lognormal background b (b is now the actual background in physical units) and A_n is not shared. The type-0 background μ_0 comes from the distribution Gamma(μ_0/T, $N_B + \alpha_0$, $T_B + \beta_0$) as already discussed.

The four-dimensional integration in Eq. (34) can be done using Monte Carlo, where A_n, A_s and b are generated from lognormal distributions and μ_0 is generated from a Gamma distribution. In order to have fast computations involving likelihood functions, for example for use in calculating Bayesian integrals, a table of values of the likelihood function is calculated numerically for each measurement, and these results are stored for use in repeated interpolations. It is important to recognise that the likelihood function is the statistical expression of the data and is independent of subsequent interpretation of the data.

When the counts are large and the lognormal distributions are narrow (S small) so that they are approximately normal, one can use the fact that, considered as a function of x,

$$x^n \exp(-x) \propto \exp\left(-\frac{1}{2n}(x-n)^2\right) \cong \exp\left(-\frac{1}{2}\left(\frac{\ln(x/n)}{1/\sqrt{n}}\right)^2\right) \tag{35}$$

for large n. Both the Poisson term and the Gamma distribution can then be approximated as normal distributions, and Eq. (34) becomes the convolution of normal distributions and is approximately

$$L(\psi) \propto \frac{1}{\sigma(\psi)} \exp\left(-\frac{1}{2}\left(\frac{y - a\psi}{\sigma(\psi)}\right)^2\right), \tag{36}$$

using Eq. (30) for $\sigma^2(\psi) = \mathrm{Var}(y)$.

Another approximation that applies when the normalisation uncertainty is not small is the lognormal

$$L(\psi) \propto \exp\left(-\frac{1}{2}\frac{(\ln(y/\psi))^2}{S_{\mathrm{tot}}^2}\right), \qquad S_{\mathrm{tot}}^2 = S^2 + (\sigma_c/y)^2 \tag{37}$$

where σ_c is the counting uncertainty of y without considering normalisation uncertainty. The likelihood function is then approximately lognormal. In comparing Eqs (36) and (37), one sees that Eq. (37) has a simpler form without the quantity a that occurs in Eq. (36). If the quantity A_0 in Eq. (12) had been chosen to be the mean rather than the median, Eq. (36) would take the simpler form.

Comparisons of these approximations with the exact likelihood calculation are shown in Figure 1.

Figure 1 demonstrates that in some cases with large numbers of detected counts, the exact likelihood and normal and lognormal approximations are all in fairly good agreement, which serves as a basic validation of the numerical methods.

Figure 2 shows the effect of adding in the logarithmic background with parameters given in Table B1.

The result shown in Figure 2 is unexpected. The standard deviation of the background for the $S = 2$ is about 200 counts, while the exact likelihood function is much narrower than this. This can be understood by remembering that the likelihood function is premised on the measurement of a certain number, 20, of counts, and if 20 counts have been measured it is not possible for the likelihood function to be appreciable at 200 counts, as it would be using the normal approximation where the background variance has been added into the denominator. Thus, the large ψ portion of the likelihood function does not move out as the background is increased, and the width of the likelihood function is misrepresented by the standard deviation of y.

Also, crucial in the Bayesian interpretation of the data in a situation where true positives are rare is the value of the likelihood function at $\psi = 0$, because the prior is then strongly peaked at $\psi = 0$. The posterior distribution of ψ can be completely changed by having an additional background. To obtain an accurate value of the likelihood at $\psi = 0$ requires the exact likelihood calculation. This is an important reason for using the exact likelihood

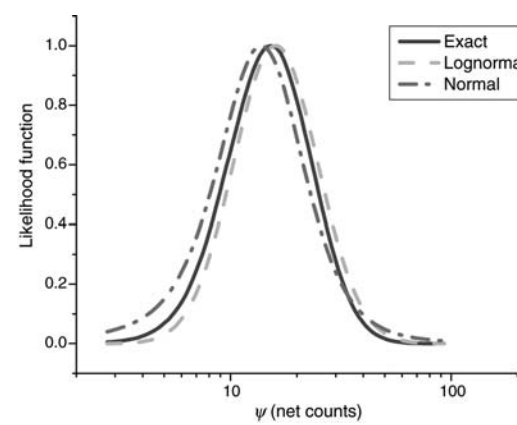

Figure 1. Likelihood function calculations for the uranium parameters as given in Table 5 assuming 20 counts have been detected, but assuming no lognormal background. The exact expression is shown together with the normal and lognormal approximations.

Downloaded from rpd.oxfordjournals.org by guest on October 4, 2010

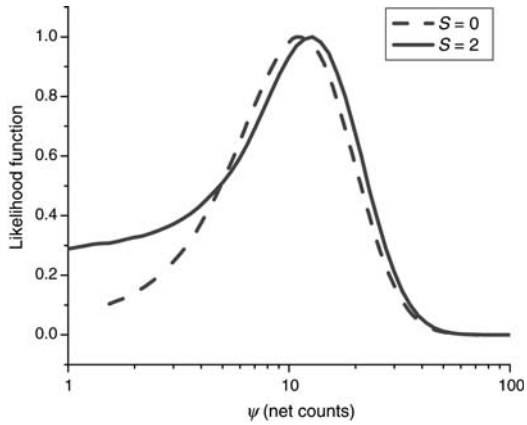

Figure 2. Exact likelihood function calculations for the uranium parameters as given in Table B1 assuming that 20 counts have been detected. An additional lognormal background is assumed with a 4-count median as given in Table B1. Two cases are shown: (1) the background has logarithmic standard deviation $S = 2$ as given in Table B1 (median 4 counts, standard deviation 238 counts) and (2) the background has logarithmic standard deviation $S = 0$ (median 4 counts, standard deviation 0).

calculations in such cases. An example involving detection of ^{234}U in urine is given in what follows.

Measures of statistical self-consistency

One can calculate χ^2 (ψ) using the posterior mean as an estimate of the true value in order to check that χ^2/N_{data} is not much larger than one and that the model is therefore not statistically inconsistent with the data. The formulas given here allow exact calculation of χ^2 (ψ) within this statistical model. However, in practice a better 'χ^2' is found to be χ^2_{like} given by two times the posterior average of the log of the likelihood function, because of slow convergence of χ^2/N_{data} with respect to N_{data} for lognormally distributed data, as discussed in Appendix C. For data that is approximately lognormally distributed, the log-transformed χ^2, similar to Eq. (37), given by

$$\chi^2_{\text{LN}} = \sum_{j=1}^{N_{\text{data}}} \frac{\left(\ln(y_j/\psi)\right)^2}{S^2_{\text{tot},j}}$$

does not have the slow convergence problem, and is more appropriate. As shown by Eq. (37), the quantity χ^2_{like} reduces to χ^2_{LN} in this limiting case. Hence, in both the cases of normally or lognormally distributed data, the quantity χ^2_{like}, obtained from the exact likelihood calculation, reproduces the appropriate expressions for χ^2 and therefore seems a reasonable general choice as a measure of statistical self-consistency.

For a visual representation of the exact likelihood function one can use the maximum, as the central point, and the width of the likelihood function, as error bars. The upper width estimate is obtained by treating the likelihood function as a probability distribution and computing the difference between the true value at the maximum and the larger true value where the probability has decreased to $\exp(-1/2)$ of the maximum, and similarly for the lower width estimate. Then, visually, one can compare two posterior distributions of the true value of the measured quantity, one, the data point and error bars, from the single data point itself at the time of the measurement and the other, a curve, showing the posterior mean for all times obtained using all the data. The width of the posterior mean using all the data is relatively small and usually need not be shown. For the modelling to be self-consistent the two posteriors should be self-consistent; that is, there should not be many points that are a large number of standard deviations from the curve.

What is the significance of all this for internal dosimetry?

In order to understand the practical significance of all this, a specific example is considered. Urine samples are taken from a person, and one would want to infer if occupationally related ^{234}U is present. As discussed in Appendix B, a lognormal environmental background of ^{234}U is assumed with a median value $=\hat{b}/A_0 = 1$ mBq and logarithmic standard deviation $S_b = 2$. Simulated urine data are generated assuming that an intake occurred on 10 March 2005, corresponding to a dose of 5 mSv, CED. The data are shown in Table 5.

The normalisation uncertainty is $S = 0.3$, the type-0 background has $\alpha_0 = 2$ and $\beta'_0 = 0.375$, and there is an environmental background as already discussed, with median value 1 mBq and $S = 2$. All the lognormal uncertainty is shared, and $S_s = 0.3$. There is a 5-mSv, type S, 5-μm AMAD inhalation intake occurring on 10 March 2005.

A data plot is shown in Figure 3.

The data uncertainty is dominated by the uncertainty of the environmental background, with a standard deviation of about 60 mBq. As seen in Figure 3, none of the data exceed even a single standard deviation, so the classical decision level method with a decision level of 1.645 standard deviations leads to a no-intake interpretation.

Figure 4 shows the data along with the calculated Bayesian interpretation. Exact likelihood function calculations were used and are essential.

The data is represented as the maximum and width of the exact likelihood as already discussed.

Downloaded from rpd.oxfordjournals.org by guest on October 4, 2010

Table 5. Simulated ^{234}U urine bioassay data.

Date	y (mBq)	σ_c (mBq)	N	N_B	f
19 July 2004	1.10	1.72	36	21	0.283
2 August 2004	0.82	1.36	44	9	0.204
16 August 2004	−6.80	0.76	8	19	0.255
30 August 2004	−6.73	1.37	18	64	0.297
13 September 2004	0.81	1.63	30	5	0.296
27 September 2004	−6.30	0.95	14	35	0.239
11 October 2004	−8.48	0.24	0	10	0.328
25 October 2004	−4.11	1.25	16	18	0.304
8 November 2004	−8.19	1.02	4	22	0.448
23 November 2004	−0.91	1.44	26	7	0.280
7 December 2004	7.77	2.59	61	57	0.323
21 December 2004	−8.15	0.19	0	5	0.368
4 January 2005	−9.29	1.41	6	40	0.497
18 January 2005	−7.20	0.81	5	15	0.335
1 February 2005	−0.34	1.80	25	18	0.354
15 February 2005	−2.40	1.55	28	40	0.283
1 March 2005	−7.51	0.61	4	14	0.280
15 March 2005	18.73	3.72	64	28	0.459
29 March 2005	−3.10	1.78	15	20	0.444
12 April 2005	0.88	1.88	37	37	0.301
26 April 2005	4.49	1.74	51	3	0.243
10 May 2005	−4.80	0.96	17	18	0.227
24 May 2005	−5.74	1.53	15	42	0.369
7 June 2005	−4.26	1.57	8	7	0.542
21 June 2005	9.08	2.28	73	40	0.263
5 July 2005	4.40	2.14	41	19	0.331
19 July 2005	−5.53	0.89	15	22	0.221
2 August 2005	−6.21	0.76	13	24	0.202
16 August 2005	−5.92	0.73	9	6	0.238
30 August 2005	−5.87	1.11	19	49	0.238
14 September 2005	−3.03	1.53	14	10	0.401
28 September 2005	−4.35	1.15	16	17	0.279
12 October 2005	32.81	3.46	142	10	0.290
26 October 2005	−4.61	1.27	12	14	0.355
9 November 2005	−5.51	0.83	12	11	0.234
23 November 2005	−4.93	1.00	14	14	0.259
7 December 2005	−5.69	1.19	10	19	0.360
21 December 2005	−0.52	2.54	32	66	0.426
4 January 2006	7.34	2.62	46	28	0.381
18 January 2006	−5.08	1.28	10	14	0.390
1 February 2006	−4.88	1.39	13	23	0.368

The measured quantity is the 24-h urine excretion. The quantity of interest y (urine 24 h) is given by $f(N - N_B/R) - b$, where b is the lognormally distributed background with median value 1 mBq and $S = 2$ (apparent average = 7.73 mBq) and $R = 4.375$. The quantity σ_c is the counting uncertainty of y without considering normalisation uncertainty. The quantities N_B and R are modified by the Bayesian replacements $N_B + \alpha_0 \rightarrow N_B$ and $R + \beta_0' \rightarrow R$ with $\alpha_0 = 2$ and $\beta_0' = 0.375$.

The upper width estimate is obtained by treating the likelihood function as a probability distribution and computing the difference between the true value at the maximum and the larger true value where the probability has decreased to $\exp(-1/2)$ of the maximum, and similarly for the lower width estimate.

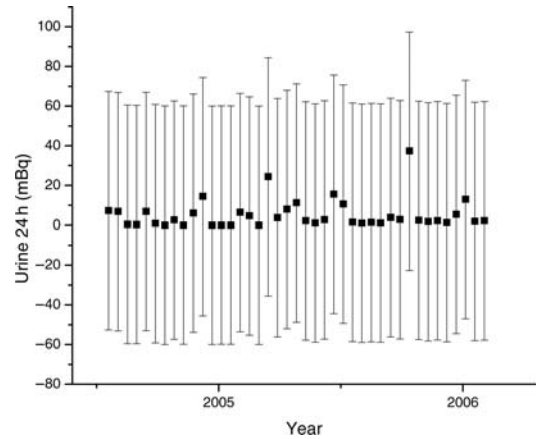

Figure 3. Simulated ^{234}U urine data. There is a single intake, on 10 March 2005, corresponding to a dose of 5 mSv, CED.

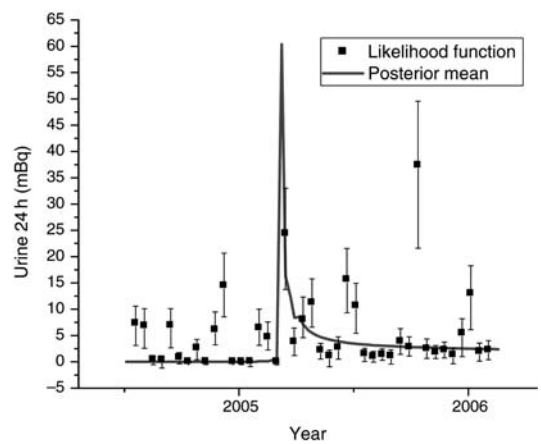

Figure 4. Simulated ^{234}U urine data and the Bayesian interpretation. The plot displays the data as the maximum and lower and upper width estimates obtained from the exact likelihood function.

The Los Alamos ID code (version IDT.0e) was used[8,9]. There were 40 (potential) intakes with variable times situated in the 40 monitoring intervals between bioassay data points. The prior probability of the date of intake was uniform within each monitoring interval. The prior probability of intake amount for each intake was given by the alpha prior[10] with a probability of intake $\alpha = 0.001$ per year, corresponding to real intakes being very rare. Nine inhalation biokinetic models were allowed: types F, M and S for 1, 5 and 10 μm AMAD, respectively. The prior probabilities of all biokinetic models are equal.

The reconstructed 2005 CED is shown in Figure 5. Two independent MCMC chains are run

Downloaded from rpd.oxfordjournals.org by guest on October 4, 2010

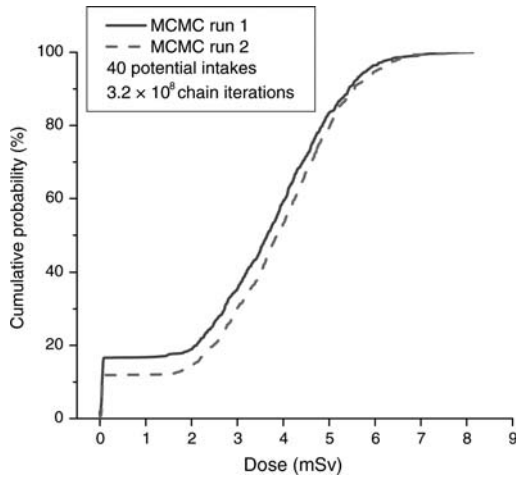

Figure 5. Posterior distribution of 2005 CED. The two curves are for different MCMC runs with different random number seeds and different starting points. The number of chain iterations for this degree of convergence was 3.2×10^8.

to check convergence, with different random number seeds and the most extreme starting points (smallest and largest values). To obtain the degree of convergence shown required about a third of a billion (3.2×10^8) iterations of the chain! The computing time was about 100 min using a 3-GHz workstation.

The probability[11] of the hypothesis that intakes have occurred is 99 %; however, as seen in Figure 5 the probability of a very small dose is about 14 %. By the hypothesis test, some occupational dose is quite certain, and the dose would be reported. The two criteria (posterior probability of two hypotheses and posterior probability of total CED) are expected theoretically to be the same when the intake prior is a mixture of a delta function at zero and another distribution separated from zero. The two criteria were found to give reasonable agreement in a study of single-intake situations.

In Bayesian hypothesis, testing the probability of a hypothesis is proportional to the average of the likelihood function over the prior, which is 2×10^{-19} for the interpretation shown in Figure 5 and 2×10^{-21} for the no-intake hypothesis. As shown in Table 6, there is a difference of five orders of magnitude between averaging over the prior and posterior. Classical hypothesis testing would compare the maximum likelihood values of χ^2, which is similar to comparing posterior average values.

Note that hypothesis testing asks which of the two hypotheses is most likely given the data. Statistical self-consistency testing is more open ended. It asks whether the collection of biokinetic models and indeed the entire formulation is adequate to represent the data.

Table 6. Numerical values of quantities used for hypothesis testing.

Hypothesis	No intakes	Intakes
Quantities		
χ^2_{like}	2.4×40	1.6×40
$\exp(-\chi^2_{\text{like}}/2)$	1.4×10^{-21}	1.3×10^{-14}
Likelihood-prior average	2×10^{-21}	2×10^{-19}
Likelihood-post average	2×10^{-21}	4×10^{-14}

By comparing the likelihood function averaged over the prior for the two hypotheses, one concludes that the probability of no intakes is about 1 % (for equal prior probability of the two hypotheses).

In this example, it is seen that detecting an occupational intake of type S uranium, even as large as 5 mSv CED, is completely impossible using the classical decision level approach, whereas using a Bayesian approach with exact likelihood calculations, the intake is detected with 99 % probability and with the posterior mean value within 30 % of the correct value.

DISCUSSION

The present work corrects an important error in a previous paper. The discussion of the Gaussian approximation to the exact likelihood function in Ref. 3 erroneously contains the expression $N_B + \alpha - 1$, where it should properly be $N_B + \alpha$. The zero background count problem is simply taken care of with the uniform-prior Bayesian approach ($\alpha = 1$, $\beta \rightarrow 0$), as shown here and elsewhere.

The Bayesian approach to statistical inference in counting measurements has been treated very well before, two notable examples being refs (12) and (13). In ref. (12), the authors derive the posterior distribution of the true counting rate given the measured number of counts without explicit reference to Bayes theorem with only the comment 'a little thought will show that . . . [the Poisson formula] is also the solution of the inverse problem' . . . That is, the authors 'little thought' has come up with the uniform-prior Bayes formula. Ref. (13) contains a full, explicitly Bayesian treatment of the problem including a good discussion of prior probability distributions. Ref. (14) contains a uniform-prior Bayesian treatment of the net counting rate with no radioactivity in the sample.

Relative to Refs. (4) and (5), the 'shared' uncertainty of a lognormal background defined here corresponds to the 'measurement' uncertainty σ_m in these papers, and the results are in agreement in regards to this background. However, the treatment of the uncertainty related to measured background

Downloaded from rpd.oxfordjournals.org by guest on October 4, 2010

counts N_B differs because of missing the contribution from Var(μ) in Eq. (6) (for this background, μ was treated as constant in these papers).

It is interesting to compare the formula for $\sigma_0^2\hat{A}^2$ in Table 2 with the work of Strom and MacLellan[15]. They state eight different formulas for the closely related classical decision level. The flat prior result here ($\alpha = 1$, $\beta \to 0$) matches the Bayesian result given by their Eq. (9), where the decision level for net counts is stated as $k_\alpha/T_B\sqrt{(N_B+1)(1/R+1/R^2)}$ with k_α found from a cumulative normal distribution with a specified upper tail area α representing the false-positive rate.

Since for small numbers of counts the distributions are far from normal, for calculation of false-positive rates with a classical decision level, it would seem better to use the Eq. (20) to obtain the probability distribution of gross counts as illustrated in Appendix A, as has been discussed by Justus[16]. This allows straightforward calculation of an integer gross-count alarm set point with a desired false-positive rate.

FUNDING

This work was funded in part by Los Alamos National Laboratory/US Department of Energy and the Russian Federal Medico-Biologic Agency.

REFERENCES

1. National Council on Radiation Protection. *Uncertainties in internal radiation dosimetry*. NCRP Report SC 6.3 (in preparation), status report available at: http://www.ncrponline.org/Current_Prog/SC_6-3.html (accessed 7 February 2009).
2. Bulmer, M. G. *On fitting the Poisson lognormal distribution to species-abundance data*. Biometrics **30**(1), 101–110 (1974).
3. Miller, G., Martz, H. F., Little, T. and Guilmette, R. *Using exact Poisson likelihood functions in Bayesian interpretation of counting measurements*. Health Phys. **83**, 512–518 (2002).
4. Miller, G. *Statistical modeling of Poisson-lognormal data*. Radiat. Prot. Dosimetry **124**(2), 155–163 (2007). Available on http://rpd.oxfordjournals.org/cgi/reprint/124/2/155 (accessed 7 February 2009).
5. Miller, G., Vostrotin, V. and Vvedensky, V. *Uncertainties of Mayak urine data*. Radiat. Prot. Dosimetry **133**, 171–176 (2009).
6. Miller, G and Inkret, W. C. '*Excretion-time normalization of urine samples using specific gravity*'. Los Alamos Radiological Dose Assessment Technical Issue Paper 011 (1999).
7. Moss, W. D., Campbell, E. E., Schulte, H. F. and Tietjen, G. L. *A study of the variations found in plutonium urinary data*. Health Phys. **17**, 571–578 (1969).
8. Miller, G., Martz, H. F., Little, T and Guilmette, R. *Bayesian internal dosimetry calculations using Markov Chain Monte Carlo*. Radiat. Prot. Dosimetry **98**, 191–198 (2002).
9. LANL. *Los Alamos Center for Bayesian Methods in Environment, Safety and Health* (2007) (www.lanl.gov/bayesian).
10. Miller, G., Inkret, W. C., Little, T. T., Martz, H. F. and Schillaci, M. E. *Bayesian prior probability distributions for internal dosimetry*. Radiat. Prot. Dosimetry **94**, 347–352 (2001).
11. Miller, G., Martz, H., Little, T. and Bertelli, L. *Bayesian hypothesis testing-use in interpretation of measurements*. Health Phys. **94**, 248–254 (2008).
12. Rainwater, L. J. and Wu, C. S. *Applications of probability theory to nuclear particle detection*. Nucleonics 60–69 October (1947).
13. Annis, M., Cheston, W. and Primakoff, H. *On statistical estimation in physics*. Rev. Mod. Phys. **25**(4), 818–830 (1953).
14. Groer, P. G. *Exact and approximate Bayesian estimation of net counting rates*. Radiat. Prot. Dosimetry **102**(3), 265–268 (2002).
15. Strom, D. J. and MacLellan, J. A. *Evaluation of eight decision rules for low-level radioactivity counting*. Health Phys. **81**(1), 27–34 (2001).
16. Justus, A. *Poisson-based decision levels and detection limits applicable to the use of health physics instrumentation*. Argonne National Laboratory Health Physics, Technical Note (2002).
17. Little, T., Miller, G., Guilmette, R. and Bertelli, L. *Uranium dose assessment: a Bayesian approach to the problem of dietary background*. Radiat. Prot. Dosimetry **127**, 333–338 (2007).

APPENDIX A—EXAMPLES OF ANALYSIS OF TYPE-0 BACKGROUND

For this example, it is assumed that three background counts have been detected with a uniform-prior probability distribution of the true background counting rate. A Fortran program that calculates the posterior distribution of counting rate using Monte Carlo is shown in Figure A1.

The Monte Carlo-generated distribution of the true counting rate is shown in Figure A2.

Assuming that α and β' are known, the distribution of counts N coming from the background (no radioactivity in the sample) would be as shown in Figure A3.

The prior parameters α and β can also be determined empirically from a large data set. For example, using the ^{234}U data already discussed for some 8000 measurements, the empirical cumulative distribution of background counts is as shown in Figure A4, where the parameters α and β were determined by trial and error, guided by the fact that the most probable value of N_B is $(\alpha - 1)T_B/\beta$.

Assuming a Gamma-function prior, the expected distribution of background counts is given by Eq. (20), the cumulative distribution of which is the curve shown in Figure A4. The calculations used the FACT function in Excel to evaluate the factorial,

Downloaded from rpd.oxfordjournals.org by guest on October 4, 2010

```
      program main
      open(1,file='test.txt')
      xran=rand(1)

1     continue
c  generate counting rate from uniform distribution from 0 to 1
      xlambda=rand(0)
c  generate number of background counts in time = 30
c   from Poisson with mean number of counts = xlambda*30
      nb=ipoisson(xlambda*30)
c  assume 3 counts have been detected
      if(nb.ne.3) go to 1
      itrial=itrial+1
c  record xlambda when 3 counts are detected
      write(1,*)itrial,xlambda
      if(itrial.lt.10000) go to 1
      stop
      end
```

Figure A1. Fortran program that calculates the distribution of true counting rate given a uniform prior and that three background counts have been observed. The function iPoisson (μ) returns an integer from a Poisson distribution with mean value μ.

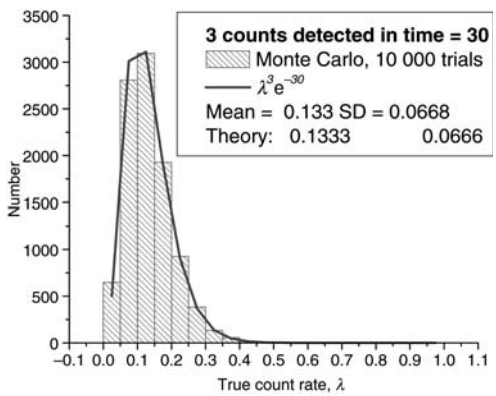

Figure A2. Monte-Carlo generated distribution of true counting rate, given that three counts have been detected. As shown, the distribution is a Gamma distribution with parameters $\alpha = N_B + 1$, $\beta = T_B$.

which allows only integer values of α, but that seems adequate. The distribution shown in Figure A4 is likely to be a mixture of several distributions that would best be separated, for example, separate distributions for each physical detector during limited periods of time around the time of the sample measurement.

The method illustrated in Figure A4 could be applied to determine the parameters α_0 and β_0 empirically.

APPENDIX B—URANIUM IN URINE

For bioassay of uranium in urine, there is an important environmental background caused by uranium in food and drinking water. Many people have uranium in their bodies from environmental intakes and always

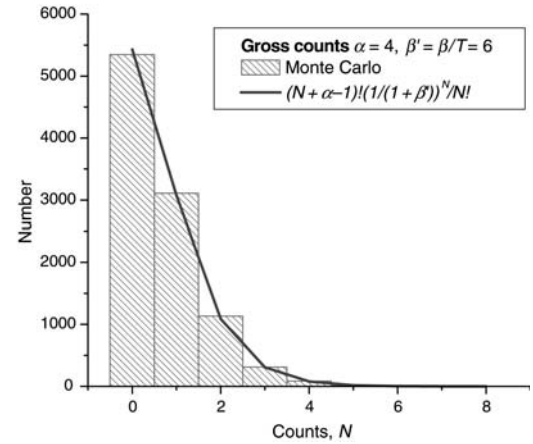

Figure A3. Monte-Carlo generated distribution of gross counts and analytical formula assuming no radioactivity in the sample and that the background counting rate is given by a Gamma distribution with $\alpha = 4$, $\beta/T = 6$. Such a curve can be used to calculate alarm set points on gross counts for a specified false-positive rate.

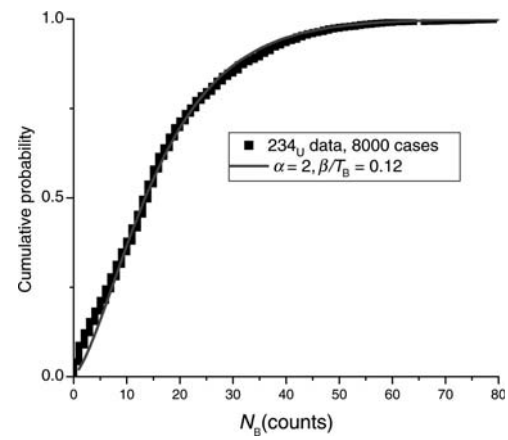

Figure A4. Empirical cumulative distribution of background counts for ^{234}U. The curve assumes a Gamma function prior to parameters as indicated.

excrete a certain amount of uranium. This amount is highly variable for a single person and also from person to person[17]. It is assumed that this background in urine results from randomly occurring dietary intake of soluble (e.g. type-F) uranium. The appendix uses data from Los Alamos to parameterise the form and magnitude of this background.

Figure B1 shows data from Los Alamos (with type-0 background subtraction) for the five individuals with the largest number of urine measurements (about 200 measurements). The number of occupational intakes is very small, and hence this

Downloaded from rpd.oxfordjournals.org by guest on October 4, 2010

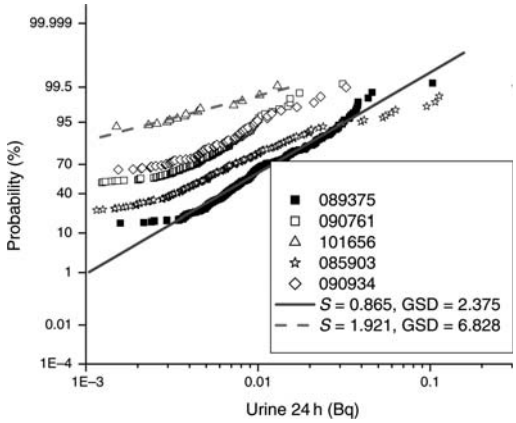

Figure B1. Observed cumulative distribution of ^{234}U activity in urine for five Los Alamos workers. Only results greater than 3 standard deviations (from counting statistics) positive are shown. The distributions are roughly lognormal with quite large geometric standard deviations (GSD).

distribution is assumed to be the result of environmental intakes only. The data that appear in the plot are at least three standard deviations positive (counting statistics only). The reason for this is so that the distribution represents the underlying distribution of uranium in urine, not confused by measurement uncertainty. Thus, the cumulative distributions begin with the percentage of results that are less than three standard deviations positive. The distributions are seen to be roughly lognormal.

For the sake of having a single, rather extreme, numerical example, it is assumed that a possible value of lognormal environmental background of ^{234}U is a median value $=\hat{b}/A_0 = 1$ mBq and logarithmic standard deviation $S_b = 2$. Other assumed parameter values are shown in Table B1.

Table B1. Numerical example of the environmental background.

Parameter	Value
A_0 (count Bq^{-1})	4000
N_B	16
R	4
α_0	2
β_0	0.375
S_x	0.3
S_r	0.05
S_c	0.05
S_{xrc}	0.308
a_{xrc}	1.049
S_b	2.000
a_b	7.389
a_{bxrc}	7.751
S_{bxrc}	2.024
$\hat{b}$ (count)	4

The variance of y is then given by

$$\begin{aligned}\mathrm{Var}(y) &= \frac{1}{A_0^2}\frac{(N_B + \alpha_0)}{R + \beta_0'}\left(1 + \frac{1}{R + \beta_0'}\right) \\ &\quad + \frac{1}{A_0^2}\hat{b}a_{bxrc}\left(1 + \hat{b}a_{bxrc}\left(a_{bxrc}^2 - 1\right)\right) \\ &\quad + \frac{a_{xrc}\psi}{A_0}\left(1 + 2\hat{b}a_{bxrc}\left(a_{xrc}^2 - 1\right)\right) \\ &\quad + (a_{xrc}\psi)^2\left(a_{xrc}^2 - 1\right) \\ \mathrm{Var}(y) &= 0.316\,\mathrm{mBq}^2 + 3550\,\mathrm{mBq}^2 + \psi 1.9\,\mathrm{mBq} \\ &\quad + \psi^2 0.11\end{aligned} \tag{A.1}$$

In estimating the lognormal standard deviation of the environmental background, note that what is obtained from background data is S_{bxrc} so that $S_b^2 = S_{bxrc}^2 - S_{xrc}^2$, and one needs to be aware of the effect of measurement uncertainties. For the zero true amount, $Var(y)$ in Eq. (A.1) is very nearly the same as the variance of the lognormal background alone ($(59.6\,\mathrm{mBq})^2 = 3550\,\mathrm{mBq}^2$). The term linear in ψ is increased by a factor of more than 6 ($=2\hat{b}a_{bxrc}(a_{xrc}^2 - 1)$) because of the lognormal background.

APPENDIX C—WHY LOGNORMAL DATA ARE PROBLEMATIC IN PRACTICE

For a lognormal background with large S, the value of χ^2/N_{data} does not converge to its limiting value until N_{data} is quite large, because only very occasionally do the large fluctuations that dominate χ^2 occur. This phenomenon is illustrated by the Monte Carlo calculation shown in Table C1.

Table C1. χ^2 calculated around the true theoretical mean value with respect to the true theoretical variance for lognormal data with $S = 2$ and $S = 1$.

N_{trials}	Average χ^2	
	$S = 2$	$S = 1$
10^1	0.008 ± 0.002	0.434 ± 0.173
10^2	0.045 ± 026	0.876 ± 0.397
10^3	0.147 ± 0.049	0.943 ± 0.188
10^4	0.233 ± 0.042	0.877 ± 0.074
10^5	0.474 ± 0.066	0.945 ± 0.034
10^6	0.719 ± 0.089	0.999 ± 0.011

For large S, this demonstrates very slow convergence and that the statistical standard deviation (shown after the $\pm$) can be misleading.

Downloaded from rpd.oxfordjournals.org by guest on October 4, 2010